

CONSISTENT BICLUSTERING

BY CHERYL J. FLYNN AND PATRICK O. PERRY

New York University

Biclustering, the process of simultaneously clustering observations and variables, is a popular and effective tool for finding structure in a high-dimensional dataset. A variety of biclustering algorithms exist, and they have been applied successfully to data sources ranging from gene expression arrays to review-website data. Currently, while biclustering appears to work well in practice, there have been no theoretical guarantees about its performance. We address this shortcoming with a theorem providing sufficient conditions for asymptotic consistency when both the number of observations and the number of variables in the dataset tend to infinity. This theorem applies to a broad range of data distributions, including Gaussian, Poisson, and Bernoulli. We demonstrate our results through a simulation study and with examples drawn from microarray analysis and collaborative filtering.

1. Introduction. Suppose we are given a data matrix $\mathbf{X} = [X_{ij}]$, and our goal is to cluster the rows and columns of \mathbf{X} into meaningful groups. For example, X_{ij} could be the log activation level of gene j in patient i ; our goal is to seek groups of patients with similar genetic profiles, while at the same time finding groups of genes with similar activation levels. Alternatively, X_{ij} can indicate whether or not user i reviewed movie j ; our goal is to simultaneously cluster the users and the movies. [Mirkin \(1996\)](#) termed the general clustering process “biclustering,” but it is also known as direct clustering ([Hartigan, 1972](#)), block modeling ([Arabie, Boorman and Levitt, 1978](#)), and co-clustering ([Dhillon, 2001](#)).

Empirical results from a broad range of disciplines indicate that biclustering data is useful in practice. For example, [Ungar and Foster \(1998\)](#) and [Hofmann \(1999\)](#) found that biclustering helps identify structure in latent class models for collaborative filtering problems where the data is sparse and diverse tastes make it difficult to cluster purely based on purchasing or review habits.

Several biclustering applications exist in the biological sciences. [Eisen et al. \(1998\)](#) was one of the first papers to note the benefits of clustering genes

AMS 2000 subject classifications: Primary 62H30; secondary 62P10, 62P25

Keywords and phrases: Biclustering, blockmodel, profile likelihood, consistency, microarray data, collaborative filtering

and conditions in microarray data, finding that genes with similar functions cluster together. Recently, [Harpaz et al. \(2010\)](#) applied biclustering methods to a Food and Drug Administration report database, identifying associations between certain active ingredients and adverse medical reactions. Several other applications of biclustering exist; see [Cheng and Church \(2000\)](#), [Getz, Levine and Domany \(2000\)](#), [Lazzeroni and Owen \(2002\)](#), and [Kluger et al. \(2003\)](#) as well as [Madeira and Oliveira \(2004\)](#) for a comprehensive survey.

Although their goals are the same, the references above use a variety of different biclustering algorithms. Clearly, many of these algorithms work well in practice, but they are often ad-hoc, and there are no rigorous guarantees as to their performance. In particular, the lack of any notion of consistency means that practitioners cannot be assured that their discoveries from biclustering will generalize or be reproducible; collecting more data may lead to completely different biclusters.

Our first objective in this report is to establish a probabilistic model for the data matrix, so that biclustering can be formalized as an estimation problem. Once we have done this, we study a class of biclustering algorithms based on profile-likelihood, as proposed by [Ungar and Foster \(1998\)](#). We show that these profile-based procedures are asymptotically consistent as the dimensions of the matrix tend to infinity, under weak assumptions on the elements of \mathbf{X} . Notably, our methods can handle both dense and sparse data matrices. To our knowledge, this is the first general consistency result for a biclustering algorithm.

Our work was inspired by recent developments in clustering methods for undirected networks. In that context, \mathbf{X} is a symmetric binary matrix, and the clusters for the rows of \mathbf{X} are the same as the clusters for the columns of \mathbf{X} . [Bickel and Chen \(2009\)](#) proposed using [Holland, Laskey and Leinhardt's \(1983\)](#) stochastic block model for the data matrix. They studied a general class of clustering algorithms and derived sufficient conditions for these algorithms to be consistent. [Choi, Wolfe and Airoldi \(2012\)](#) generalized this result by allowing the number of blocks to increase with the network size, [Zhao, Levina and Zhu \(2011a\)](#) allowed for nodes that do not belong to any community, and [Zhao, Levina and Zhu \(2011b\)](#) relaxed the assumption of stochastic equivalence within each block by incorporating individual effects. [Rohe, Chatterjee and Yu \(2011\)](#) also studied the stochastic block model with an increasing number of blocks and established a bound on the number of misclassified nodes for spectral clustering.

In the setting of biclustering for directed networks [Rohe and Yu \(2012\)](#) considered a stochastic block model and proposed the “dice ’em” biclustering algorithm (DI-SIM), which “dices” the rows and columns of a matrix

into clusters using its left and right singular vectors. They derived consistency results for this algorithm but required the expected degree to grow sufficiently fast. Our theoretical results do not require this restriction and are not limited to directed networks.

The remainder of the paper is organized as follows. Section 2 describes the block model and defines the profile-likelihood. Section 3 contains the main theoretical results and Section 4 presents the main steps of the proof. Section 5 corroborates the theoretical findings through the use of asymptotic simulations and the results suggest that the profile-likelihood outperforms other biclustering methods. Section 6 looks at applications to real data and the results from biclustering experiments done on a movie-review and a microarray dataset are reported. Section 7 presents some concluding remarks and the appendix includes additional technical and empirical results.

2. Block models and profile likelihood. As above, let $\mathbf{X} = [X_{ij}] \in \mathbb{R}^{m \times n}$ be a data matrix. One way to formalize the biclustering problem is to posit the existence of K *row classes* and L *column classes*, such that the mean value of entry X_{ij} is determined solely by the classes of row i and column j . That is, there is an unknown row class membership vector $\mathbf{c} \in K^m$, an unknown column class membership vector $\mathbf{d} \in L^n$, and an unknown mean matrix $\mathbf{M} = [\mu_{kl}] \in \mathbb{R}^{K \times L}$ such that

$$\mathbb{E} X_{ij} = \mu_{c_i d_j};$$

we refer to this model as a *block model*, after the related model for undirected networks proposed by Holland, Laskey and Leinhardt (1983). With a block model, our goal is to estimate \mathbf{c} and \mathbf{d} . We do this by assigning labels to the rows and columns of \mathbf{X} , codified in vectors $\mathbf{g} \in K^m$ and $\mathbf{h} \in L^n$. Ideally, \mathbf{g} and \mathbf{h} match \mathbf{c} and \mathbf{d} . Note that we are assuming that the true numbers of row and column clusters, K and L , are fixed and known.

We can employ profile-likelihood (Murphy and van der Vaart, 2000) for the biclustering task. Initially, we propose a simple distribution for \mathbf{X} , and we derive the maximum profile likelihood estimator for \mathbf{c} and \mathbf{d} under this model. Later, we consider a far more general distribution for \mathbf{X} , and we show that the simple profile-likelihood based estimator is still consistent, even under model misspecification.

Suppose that the elements of \mathbf{X} are independent draws whose distribution is specified by a single-parameter exponential family. Conditional on \mathbf{c} and \mathbf{d} , entry X_{ij} has density $g(x; \eta_{c_i d_j})$ with respect to some σ -finite measure ν , where

$$g(x; \eta) = \exp\{x\eta - \psi(\eta)\};$$

$\psi(\eta)$ is the cumulant function, and $\eta_{kl} = (\psi')^{-1}(\mu_{kl})$. With labels \mathbf{g} and \mathbf{h} , the complete data log-likelihood is

$$l(\mathbf{g}, \mathbf{h}, \mathbf{M}) = mn \sum_{k,l} \hat{p}_k \hat{q}_l \{ \bar{X}_{kl} \eta_{kl} - \psi(\eta_{kl}) \},$$

where

$$\hat{p}_k = \frac{1}{m} \sum_i \mathbf{I}(g_i = k), \quad \hat{q}_l = \frac{1}{n} \sum_j \mathbf{I}(h_j = l),$$

and

$$\bar{X}_{kl} = \frac{\sum_{i,j} X_{ij} \mathbf{I}(g_i = k, h_j = l)}{\sum_{i,j} \mathbf{I}(g_i = k, h_j = l)}.$$

The profile log-likelihood is

$$\text{pl}(\mathbf{g}, \mathbf{h}) = \sup_{\mathbf{M}} l(\mathbf{g}, \mathbf{h}, \mathbf{M}) = mn \sum_{k,l} \hat{p}_k \hat{q}_l \psi^*(\bar{X}_{kl}),$$

where $\psi^*(x) = \sup_{\eta} \{x\eta - \psi(\eta)\}$ is the convex conjugate of ψ , known as the *rate function* in the large deviations literature (Dembo and Zeitouni, 1998). Following the above derivation, it is natural to estimate \mathbf{c} and \mathbf{d} by the label vectors which maximize $\text{pl}(\mathbf{g}, \mathbf{h})$.

In the sequel, we consider a far more general setting. We consider criterion functions of the form

$$(2.1) \quad F(\mathbf{g}, \mathbf{h}) = mn \sum_{kl} \hat{p}_k \hat{q}_l f(\bar{X}_{kl}/\rho),$$

where f is any smooth convex function and ρ is a scale parameter. Borrowing terminology from the large deviations literature, we refer to f as the rate function of the criterion; though, since we allow f to take negative values, we do not require that f be a rate function in the strictest sense. We permit the elements of \mathbf{X} to have different distributions, allowing for heteroscedasticity and model misspecification. We show that under mild technical conditions, the maximizer of F is a consistent estimator of the true row and column classes.

3. Consistency results. To prove consistency, we need to work with a sequence \mathbf{X}_n of data matrices. Suppose that $\mathbf{X}_n \in \mathbb{R}^{m \times n}$ and $m = m(n)$ with $n/m \rightarrow \gamma$ for some finite constant γ . Suppose that for each n there exists a row class membership vector $\mathbf{c}_n \in K^m$ and a column class membership vector $\mathbf{d}_n \in L^n$. We assume that the components of \mathbf{c}_n are assigned independently by repeatedly drawing from a multinomial distribution with

class probability vector $\mathbf{p} \in \mathbb{R}^K$; similarly, the components of \mathbf{d}_n are assigned independently by repeatedly drawing from a multinomial distribution with class probability vector $\mathbf{q} \in \mathbb{R}^L$. When there is no ambiguity, we omit the subscript n .

Assume that the mean of element X_{ij} depends only on the row and column memberships c_i and d_j , so that

$$\mathbb{E}(X_{ij} \mid \mathbf{c}, \mathbf{d}) = \mu_{c_i d_j}$$

for some matrix $\mathbf{M} = [\mu_{kl}] \in \mathbb{R}^{K \times L}$, possibly varying with n . To model sparsity in \mathbf{X} , we allow \mathbf{M} to tend to $\mathbf{0}$. To avoid degeneracy, we suppose that there exists a sequence ρ and a fixed matrix $\mathbf{M}_0 \in \mathbb{R}^{K \times L}$ such that $\rho^{-1} \mathbf{M} \rightarrow \mathbf{M}_0$.

Define the normalized confusion matrices $\mathbf{C}(\mathbf{g}) \in \mathbb{R}^{K \times K}$ and $\mathbf{D}(\mathbf{h}) \in \mathbb{R}^{L \times L}$ by

$$C_{ak}(\mathbf{g}) = \frac{1}{m} \sum_i I(c_i = a, g_i = k), \quad D_{bl}(\mathbf{h}) = \frac{1}{n} \sum_j I(d_j = b, h_j = l).$$

Entry C_{ak} is the proportion of nodes with class a and label k ; entry D_{bl} is defined similarly. These matrices are normalized so that $\mathbf{C}^T \mathbf{1} = \hat{\mathbf{p}}$ and $\mathbf{D}^T \mathbf{1} = \hat{\mathbf{q}}$.

We only consider nontrivial partitions; to this end, for $\varepsilon > 0$, define

$$\mathcal{J}_\varepsilon = \{\mathbf{g}, \mathbf{h} : \hat{p}_k(\mathbf{g}) > \varepsilon, \hat{q}_l(\mathbf{h}) > \varepsilon\}.$$

For fixed convex rate function f , we let F be a criterion function as in (2.1).

3.1. Assumptions. Denote by \mathcal{M}_0 the convex hull of the entries of \mathbf{M}_0 . Let \mathcal{M} be a neighborhood of \mathcal{M}_0 . We require the following assumptions:

- (A1) **The biclusters are identifiable.** No two rows of \mathbf{M}_0 are equal, and no two columns of \mathbf{M}_0 are equal.
- (A2) **The rate function is locally strictly convex.** That is, $f''(\mu) > 0$ when $\mu \in \mathcal{M}$.
- (A3) **The third derivative of the rate function is locally bounded.** That is, $|f'''(\mu)|$ is bounded when $\mu \in \mathcal{M}$.
- (A4) **The average variance of the elements is of the same order as ρ .**

$$\limsup_{n \rightarrow \infty} \frac{1}{\rho m n} \sum_{i,j} \mathbb{E}[(X_{ij} - \mu_{c_i d_j})^2 \mid \mathbf{c}, \mathbf{d}] < \infty.$$

(A5) **The mean matrix does not converge to zero too quickly.** That is, $\limsup_{n \rightarrow \infty} \rho n = \infty$.

(A6) **The elements satisfy a Lindeberg condition.** For all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{\rho^2 mn} \sum_{i,j} \mathbb{E}[(X_{ij} - \mu_{c_i d_j})^2 \mathbb{I}(|X_{ij} - \mu_{c_i d_j}| > \varepsilon \sqrt{mn} \rho) \mid \mathbf{c}, \mathbf{d}] = 0.$$

A variant Lyapunov's condition (Varadhan, 2001) implies (A6). That is, if

$$\lim_{n \rightarrow \infty} \frac{1}{(\rho \sqrt{mn})^{2+\delta}} \sum_{i,j} \mathbb{E} |X_{ij} - \mu_{c_i d_j}|^{2+\delta} = 0$$

for some $\delta > 0$, then (A6) follows. In particular, for dense data (ρ bounded away from zero), uniformly bounded $(2 + \delta)$ absolute central moments for any $\delta > 0$ is sufficient. For sparse data (Bernoulli or Poisson data with ρ converging to zero), (A5) is a sufficient condition for (A6).

3.2. Main result. Given the above setup and assumptions, it follows that the maximizer of $F(\mathbf{g}, \mathbf{h})$ is a consistent estimator of the true row and column labels.

THEOREM 3.1. *Fix any $\varepsilon > 0$ with $\varepsilon < \min_a \{p_a\}$ and $\varepsilon < \min_b \{q_b\}$. Let $(\hat{\mathbf{g}}, \hat{\mathbf{h}})$ satisfy $F(\hat{\mathbf{g}}, \hat{\mathbf{h}}) = \max_{\mathcal{J}_\varepsilon} F(\mathbf{g}, \mathbf{h})$. If assumptions (A1)–(A6) hold, then all limit points of $\mathbf{C}(\hat{\mathbf{g}})$ and $\mathbf{D}(\hat{\mathbf{h}})$ are permutations of diagonal matrices, i.e. the proportions of mislabeled rows and columns converge to zero in probability.*

The statement of the theorem is somewhat awkward, because we can permute the row or column labels and get the same value of the criterion function, F . Thus, F has multiple maximizers, and it is only possible to recover the true classes up to a permutation of labels.

From the discussion in Section 2, it follows that maximizing the profile log-likelihood associated with a single-parameter exponential family may be a reasonable biclustering procedure. Furthermore, the proof of this result does not require the distribution of the data to be correctly specified and allows for possible model misspecification so long as (A1)–(A6) are satisfied. This implies that this result can be applied to binary matrices, count data, and continuous data which could be reasonably modeled by Binomial, Poisson, and Gaussian distributed data, respectively, making this a suitable result for our motivating examples.

To prove the theorem, we first establish that in the limit, F is close to a nonrandom “population version,” G . Then, we establish that G is maximized

at the true class labels. Finally, we show that outside any neighborhood around the true class labels, G is smaller than at the true values. Section 4 provides the details.

4. Proof of consistency theorem. In this section, we continue with the setup and notation of Section 3. To prove Theorem 3.1, we need to introduce some additional notation.

Define expectation matrix $\mathbf{E}(\mathbf{g}, \mathbf{h}) \in \mathbb{R}^{K \times L}$ with

$$E_{kl}(\mathbf{g}, \mathbf{h}) = \mathbb{E}(\bar{X}_{kl}(\mathbf{g}, \mathbf{h}) \mid \mathbf{c}, \mathbf{d}) = \frac{[\mathbf{C}^T \mathbf{M} \mathbf{D}]_{kl}}{[\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l},$$

where $\mathbf{C} = \mathbf{C}(\mathbf{g})$ and $\mathbf{D} = \mathbf{D}(\mathbf{h})$. Also, define normalized residual matrix $\mathbf{R}(\mathbf{g}, \mathbf{h}) \in \mathbb{R}^{K \times L}$ by

$$\mathbf{R}(\mathbf{g}, \mathbf{h}) = \rho^{-1} \{ \bar{\mathbf{X}}(\mathbf{g}, \mathbf{h}) - \mathbf{E}(\mathbf{g}, \mathbf{h}) \}.$$

The weak law of large numbers establishes that for fixed \mathbf{g} and \mathbf{h} , the convergence $R_{kl}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} 0$ holds. We can prove a stronger result, that this convergence is uniform over all \mathbf{g} and \mathbf{h} .

LEMMA 4.1. *Under assumptions (A1)–(A6), for all $\varepsilon > 0$,*

$$\sup_{\mathcal{J}_\varepsilon} \|\mathbf{R}(\mathbf{g}, \mathbf{h})\|_\infty \xrightarrow{P} 0,$$

where $\|\mathbf{A}\|_\infty = \max_{k,l} |A_{kl}|$ for any matrix \mathbf{A} .

With Lemma 4.1 (proved in Appendix A), we can establish that in the limit, $F(\mathbf{g}, \mathbf{h})$ is close to its “population version,” which depends only on \mathbf{C} and \mathbf{D} . To define this population version, first, for each M , define \mathcal{S}_M to be the set of $M \times M$ matrices with nonnegative entries summing to one: $\mathcal{S}_M = \{\mathbf{X} \in \mathbb{R}_+^{M \times M} : \mathbf{1}^T \mathbf{X} \mathbf{1} = 1\}$. Next, define function $G_{M_0} : \mathcal{S}_K \times \mathcal{S}_L \rightarrow \mathbb{R}$ to be the population version of F :

$$G_{M_0}(\mathbf{C}, \mathbf{D}) = \sum_{k,l} [\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l f\left(\frac{[\mathbf{C}^T \mathbf{M}_0 \mathbf{D}]_{kl}}{[\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l}\right).$$

LEMMA 4.2. *F is close to its population version in the sense that, for all $\varepsilon > 0$,*

$$\sup_{\mathcal{J}_\varepsilon} |F(\mathbf{g}, \mathbf{h}) - G_{M_0}(\mathbf{C}, \mathbf{D})| \xrightarrow{P} 0.$$

This is a direct consequence of Lemma 4.1 and assumption (A3), which implies that f is locally Lipschitz; see Appendix A for details.

Once we have established that F is close to its population version, our next task is to show that the population version is maximized at the true class labels.

LEMMA 4.3. *For $\delta > 0$ define*

$$\mathcal{P} = \{\mathbf{C} \in \mathcal{S}_K : \max_{a \neq a'} C_{ak} C_{a'k} < \delta\}$$

and

$$\mathcal{Q} = \{\mathbf{D} \in \mathcal{S}_L : \max_{b \neq b'} D_{bl} D_{b'l} < \delta\}.$$

If $\min_a \{[\mathbf{C}\mathbf{1}]_a\} > \eta$, $\min_b \{[\mathbf{D}\mathbf{1}]_b\} > \eta$, and $(\mathbf{C}, \mathbf{D}) \notin \mathcal{P} \times \mathcal{Q}$, then $G_{\mathbf{M}_0}(\mathbf{C}, \mathbf{D})$ is small, in the sense that

$$G_{\mathbf{M}_0}(\mathbf{C}, \mathbf{D}) - \sum_{a,b} [\mathbf{C}\mathbf{1}]_a [\mathbf{D}\mathbf{1}]_b f([\mathbf{M}_0]_{ab}) \leq -\kappa \eta^2 \delta,$$

where κ is a constant independent of δ and η .

We prove Lemma 4.3 in Appendix A. Next, we prove the consistency theorem.

PROOF OF THEOREM 3.1. Fix $\delta > 0$ and define \mathcal{P} and \mathcal{Q} as in Lemma 4.3. We will show that if $(\mathbf{g}, \mathbf{h}) \in \mathcal{J}_\varepsilon$ and if $(\mathbf{C}(\mathbf{g}), \mathbf{D}(\mathbf{h})) \notin (\mathcal{P}, \mathcal{Q})$, then $F(\mathbf{g}, \mathbf{h}) < F(\mathbf{c}, \mathbf{d})$ with probability approaching one. Moreover, this inequality holds uniformly over all such choices of (\mathbf{g}, \mathbf{h}) . Since δ is arbitrary, this implies that $\mathbf{C}(\hat{\mathbf{g}})$ and $\mathbf{D}(\hat{\mathbf{h}})$ converge to permutations of diagonal matrices, i.e. the proportions of misclassified rows and columns converge to zero.

Set $r_n = \sup_{\mathcal{J}_\varepsilon} |F(\mathbf{g}, \mathbf{h}) - G_{\mathbf{M}_0}(\mathbf{C}(\mathbf{g}), \mathbf{D}(\mathbf{h}))|$. Suppose $(\mathbf{g}, \mathbf{h}) \in \mathcal{J}_\varepsilon$. In this case,

$$\begin{aligned} F(\mathbf{g}, \mathbf{h}) - F(\mathbf{c}, \mathbf{d}) &\leq 2r_n + \{G_{\mathbf{M}_0}(\mathbf{C}(\mathbf{g}), \mathbf{D}(\mathbf{h})) - G_{\mathbf{M}_0}(\mathbf{C}(\mathbf{c}), \mathbf{D}(\mathbf{d}))\} \\ &= 2r_n + \{G_{\mathbf{M}_0}(\mathbf{C}(\mathbf{g}), \mathbf{D}(\mathbf{h})) - \sum_{a,b} [\mathbf{C}\mathbf{1}]_a [\mathbf{D}\mathbf{1}]_b f([\mathbf{M}_0]_{ab})\}. \end{aligned}$$

Pick $\eta > 0$ smaller than $\min_a \{p_a\}$ and $\min_b \{q_b\}$. By assumption, the true row and column classes follow multinomial distributions with probabilities \mathbf{p} and \mathbf{q} . Thus, for all $\mathbf{g} \in K^m$ and $\mathbf{h} \in L^n$, by the weak law of large numbers, $[\mathbf{C}(\mathbf{g})]_a \geq \eta$ and $[\mathbf{D}(\mathbf{h})]_b \geq \eta$ with probability tending to one as n increases; moreover, this holds uniformly over all choices of (\mathbf{g}, \mathbf{h}) .

Applying Lemma 4.3, to the second term in the inequality, we get that with probability approaching one,

$$F(\mathbf{g}, \mathbf{h}) - F(\mathbf{c}, \mathbf{d}) \leq 2r_n - \kappa\eta^2\delta$$

for all $(\mathbf{g}, \mathbf{h}) \in \mathcal{J}_\varepsilon$ such that $(\mathbf{C}(\mathbf{g}), \mathbf{D}(\mathbf{h})) \notin \mathcal{P} \times \mathcal{Q}$. By Lemma 4.2, $r_n \xrightarrow{P} 0$. Thus, with probability approaching one, $(\mathbf{C}(\hat{\mathbf{g}}), \mathbf{D}(\hat{\mathbf{h}})) \in \mathcal{P} \times \mathcal{Q}$. Since this result holds for all δ , all limit points of $\mathbf{C}(\hat{\mathbf{g}})$ and $\mathbf{D}(\hat{\mathbf{h}})$ must be permutations of diagonal matrices. \square

5. Empirical evaluation. We study the performance of profile log-likelihood for biclustering Bernoulli, Poisson, and Gaussian data. For these three cases, the rate functions are as follow:

$$(5.1a) \quad f_{\text{Bernoulli}}(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu),$$

$$(5.1b) \quad f_{\text{Poisson}}(\mu) = \mu \log \mu - \mu,$$

$$(5.1c) \quad f_{\text{Gaussian}}(\mu) = \mu^2/2.$$

We use the appropriate rate function depending on the context.

For the Poisson and Gaussian rate functions (5.1b) and (5.1c), the maximizer of the criterion function (2.1) does not depend on the scale factor ρ . For the Binomial rate function (5.1a), we use $\rho = 1$ in the fitting procedure, regardless of the true scale factor for the mean matrix \mathbf{M} . Even though in the simulations assumption (A1) doesn't hold for this choice of ρ , we still get consistency, because the maximizer of the criterion with $f_{\text{Bernoulli}}(\mu)$ is close to the maximizer with $f_{\text{Poisson}}(\mu)$. See Perry and Wolfe (2012) for discussion of a related phenomenon.

To maximize the profile log-likelihood, we compute initial partitions for the rows and columns by applying k -means separately to the rows and the columns. Then, we iteratively update the cluster assignments in a greedy manner, based on the Kernighan-Lin heuristic (Kernighan and Lin, 1970) employed by Newman (2006):

1. For each row and column, we compute the optimal label assignment while keeping the labels of all other rows and columns fixed; we also record the improvement made by making this assignment.
2. In order of the local improvements recorded in step 1, we perform the label reassignments determined in step 1. Note that these assignments are no longer locally optimal since the labels of many of the rows and columns change during this step.
3. Out of those labels considered in step 2, we choose the one which has the highest profile likelihood.

We iteratively perform steps 1–3 until the profile likelihood converges. The algorithm is not guaranteed to converge to a global optimum, but it seems to perform well in practice.

In our simulations, we report the proportion of misclassified rows and columns by the profile-likelihood-based method (PL), which Theorem 3.1 guarantees to be consistent. We also report misclassification errors for k -means applied separately to the rows and columns (KM) and for the DI-SIM biclustering algorithm (DS) of Rohe and Yu (2012).

For the Poisson simulation, we simulate from a block model with $K = 2$ row clusters and $L = 3$ column clusters. We vary the number of columns, n , between 200 to 1500 and we take the number of rows as $m = \gamma n$ where $\gamma \in \{0.5, 1, 2\}$.

We set the row and column class membership probabilities as $\mathbf{p} = (0.3, 0.7)$ and $\mathbf{q} = (0.2, 0.3, 0.5)$. We choose the matrix of block parameters to be

$$\mathbf{M} = [\mu_{ab}] = \frac{b}{\sqrt{n}} \begin{pmatrix} 0.92 & 0.77 & 1.66 \\ 0.17 & 1.41 & 1.45 \end{pmatrix},$$

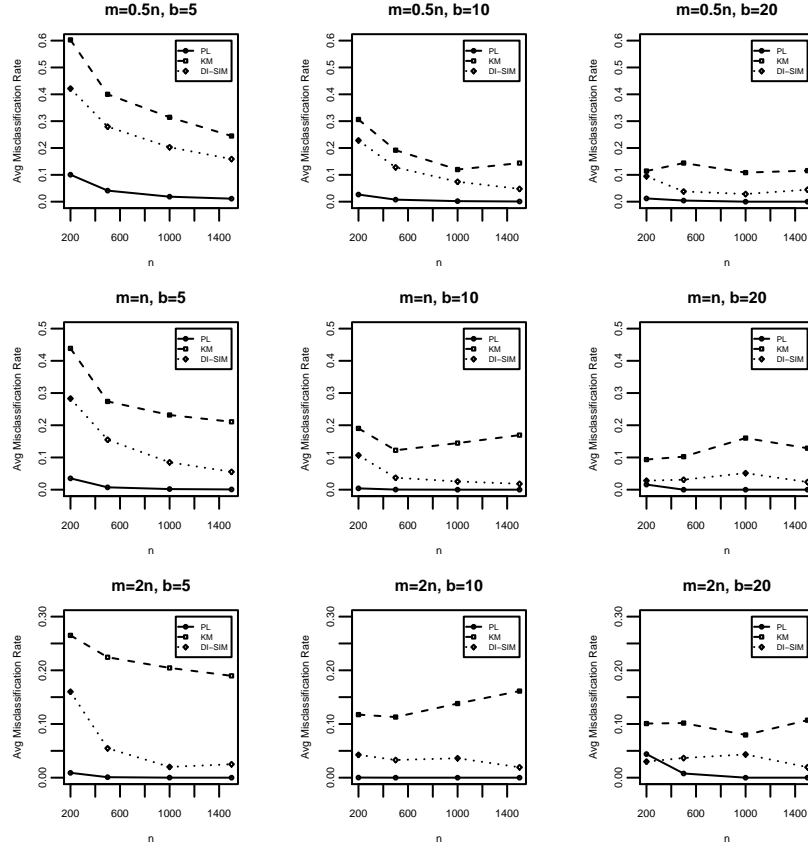
where b is chosen between 5 and 20; the entries of the matrix were chosen randomly, uniformly on the interval $[0, 2]$. Note that $\mathbf{M} \rightarrow \mathbf{0}$, so the data matrix is sparse. We generate the data conditional on the row and column classes as $X_{ij} \mid \mathbf{c}, \mathbf{d} \sim \text{Poisson}(\mu_{c_i d_j})$.

Figure 1 presents the average bicluster misclassification rates for each sample size and Table 1 reports the standard deviations. In all of the scenarios considered, the biclustering based on the profile log-likelihood criterion performs at least as well as the other methods and shows signs of convergence. Although the k -means algorithm and the DI-SIM algorithm perform well in some settings, in other settings the performance is poor and does not show signs of convergence. We also see that the standard deviation of the misclassification rate tends to zero for the PL biclustering, but not for the other two algorithms.

Appendix B describes in detail the simulations for Bernoulli and Gaussian data. For the Bernoulli simulation, we see similar behavior to the Poisson simulation reported here. For the Gaussian data, the DI-SIM algorithm is much more competitive, but our algorithm still beats it in almost all cases.

Overall, the simulation results corroborate the conclusions of Theorem 3.1 and support the use of biclustering based on the profile log-likelihood criterion.

FIG 1. Average misclassification rates for Poisson example over 100 simulations.

TABLE 1
Standard deviations for Poisson example over 100 simulations.

$m = 0.5n$								
n	PL			KM			DS	
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.0683	0.0177	0.0554	0.1007	0.1288	0.1180	0.0756	0.0586
500	0.0112	0.0046	0.0392	0.1136	0.1546	0.1668	0.0341	0.0230
1000	0.0050	0.0013	0.0002	0.1500	0.1436	0.1581	0.0222	0.0142
1500	0.0034	0.0007	0.0001	0.1613	0.1629	0.1624	0.0188	0.0089

$m = n$								
n	PL			KM			DS	
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.0158	0.0041	0.0777	0.1356	0.1444	0.1409	0.0592	0.0299
500	0.0039	0.0008	0.0000	0.1444	0.1497	0.1556	0.0260	0.0105
1000	0.0015	0.0002	0.0000	0.1619	0.1652	0.1720	0.0138	0.0065
1500	0.0007	0.0000	0.0000	0.1648	0.1672	0.1658	0.0090	0.0038

$m = 2n$								
n	PL			KM			DS	
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.0063	0.0009	0.1259	0.1182	0.1543	0.1560	0.0456	0.0636
500	0.0012	0.0000	0.0552	0.1504	0.1550	0.1569	0.0126	0.1018
1000	0.0002	0.0001	0.0000	0.1550	0.1632	0.1428	0.0052	0.1130
1500	0.0000	0.0000	0.0000	0.1578	0.1658	0.1572	0.0738	0.0848

6. Applications. In this section we use profile-likelihood-based biclustering to reveal structure in two high-dimensional datasets. For each example, we maximize the profile log-likelihood using the algorithm described in Section 5.

6.1. *MovieLens.* Collaborative filtering uses the behavior of similar consumers to provide better recommendations of products to new individuals. Since consumer habits likely vary depending on products, biclustering can help identify structure in the data and identify groups of consumers and groups of products with similar patterns. As an application of this we apply biclustering to the MovieLens data set ([GroupLens Research Project, 1998](#)).

The MovieLens data set consists of 100,000 movie reviews on 1682 movies by 943 users. Each user has rated at least 20 movies and each movie is rated on a scale from one to five. In addition to the review rating, the release date and genre of each movie is available as well as some demographic information about each user including gender, age, occupation and zip code.

In order to retain customers, movie-renting services strive to recommend new movies to individuals who are likely to view them. Since most users only review movies that they have already seen, we can use the structure of the user-movie review matrix to identify associations between users and viewing habits of movies. Specifically, we consider the 943×1682 binary matrix \mathbf{X} where $X_{ij} = 1$ if user i has rated movie j and $X_{ij} = 0$ otherwise. To find structure in \mathbf{X} , we biclustered the rows and columns of \mathbf{X} using the profile log-likelihood based on the Bernoulli criterion (5.1a).

To determine a reasonable selection for the number of biclusters we varied the number of user groups from two to three and the number of movie groups from two to seven. Qualitatively, the model with three user groups and six movie groups seems to provide a parsimonious description of the data.

Figure 2 presents the heatmap of the data based on the resulting bicluster assignments, with the ordering of the clusters determined by the total number of a reviews in each cluster. Table 2 reports the top ten movies in each group. The eclectic mix of genres within each movie group suggests that the rating behavior of users is not explained by genre alone. Figure 6.1 presents a boxplot comparing the distributions of the movie release years for each group. We can see a clear ordering of the movie groups by median release date.

The median ages within the user group were 32, 31, and 30.5, and the percentages of male users within each group were 67.6%, 72.8%, and 77.8%. These statistics suggest that there is some age and gender effect on the reviewing habits of the users.

Roughly speaking, user group 3 is consistently active across all movie groups with increasing activity as the popularity of the movie increases. The reviewing habits of user group 2 follow a similar pattern but to a lesser extent. In contrast, user group 1 is consistently inactive with the only exceptions being movie groups 4 and 6. From Figure 6.1, it appears that the users in group 1 only rate recent movies whereas users in groups 2 and 3 rate movies from all time periods.

The biclusters discovered here suggest that a movie-renting service should recommend under-reviewed movies to individuals in user group 3, and it should recommend new releases to users in group 1. This information can be used in an ensemble-based recommendation engine like that of Töschner, Jahrer and Bell (2009).

FIG 2. Heatmap generated from MovieLens data reflecting the varying review patterns in the different biclusters. Blue identifies movies with no review and white identifies rated movies.

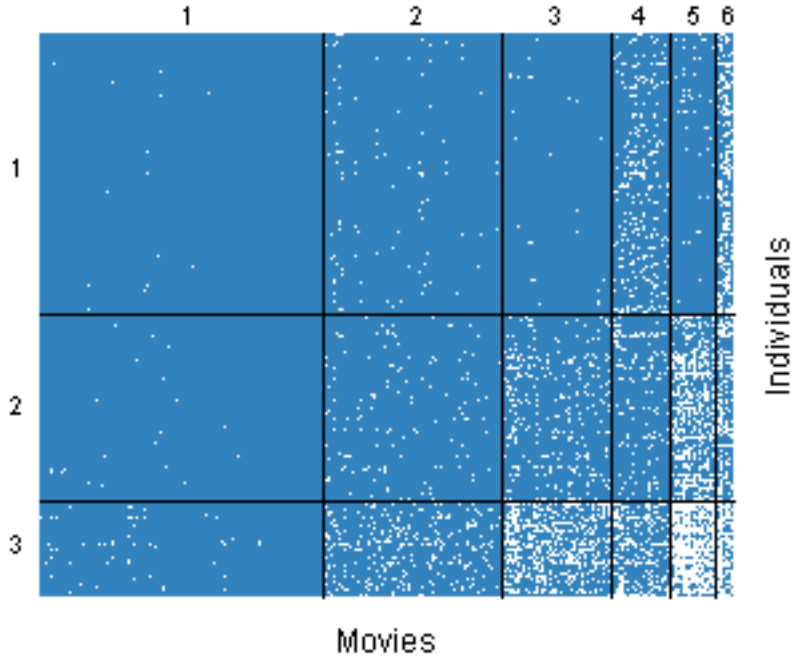
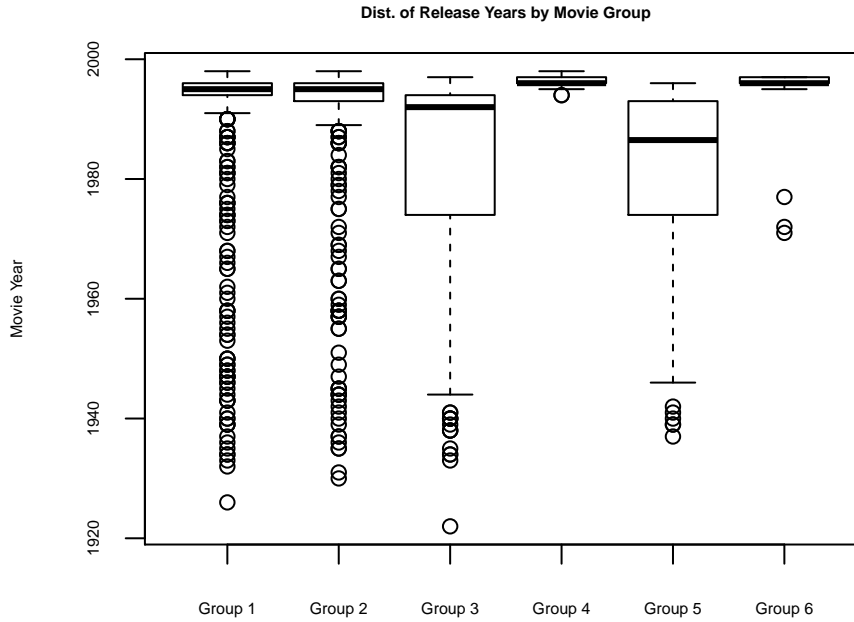


TABLE 2
The top ten movies in each cluster based on the total number of reviews.

Group 1	Group 2
Jade (1995)	She's the One (1996)
When the Cats Away (1996)	Jack (1996)
Jaws 3-D (1983)	The Preacher's Wife (1996)
Bastard Out of Carolina (1996)	Striptease (1996)
Exit to Eden (1994)	Mirror Has Two Faces, The (1996)
The Ruling Class (1972)	Hercules (1997)
The Air Up There (1994)	Kids in the Hall: Brain Candy (1996)
Bad Taste (1987)	Jean de Florette (1986)
Stuart Saves His Family (1995)	The Fan (1996)
Cabin Boy (1994)	Extreme Measures (1996)
Group 3	Group 4
The Firm (1993)	Courage Under Fire (1996)
The Abyss (1989)	Volcano (1997)
Die Hard: With a Vengeance (1995)	Murder at 1600 (1997)
Remains of the Day, The (1993)	Mars Attacks! (1996)
Sneakers (1992)	The People vs. Larry Flynt (1996)
The Professional (1994)	Starship Troopers (1997)
Clerks (1994)	Eraser (1996)
Reservoir Dogs (1992)	Das Boot (1981)
Like Water For Chocolate (1992)	Good Will Hunting (1997)
Chinatown (1974)	The Fifth Element (1997)
Group 5	Group 6
Raiders of the Lost Ark (1981)	Star Wars (1977)
Pulp Fiction (1994)	Contact (1997)
The Silence of the Lambs (1991)	Fargo (1996)
The Empire Strikes Back (1980)	Return of the Jedi (1983)
Back to the Future (1985)	Liar Liar (1997)
The Fugitive (1993)	The English Patient (1996)
Indiana Jones and the Last Crusade (1989)	Scream (1996)
The Princess Bride (1987)	Toy Story (1995)
Forrest Gump (1994)	Air Force One (1997)
Monty Python and the Holy Grail (1974)	Independence Day (1996)

FIG 3. Boxplot comparing the different clusters based on the number of movies reviewed by each individual.



6.2. *AGEMAP*. Biclustering is commonly used for microarray data to visualize the activation patterns of thousands of genes simultaneously. It is used to identify patterns and discover distinguishing properties between genes and individuals. We use the AGEMAP dataset (Zahn et al., 2007) to illustrate this process.

AGEMAP is a large microarray data set containing the log expression levels for 40 mice across 8932 genes measured on 16 different tissue types. For this analysis, we restrict attention to two tissue types: cerebellum and cerebrum. The 40 mice in the dataset belong to four age groups, with five males and five females in each group. One of the mice is missing data for the cerebrum tissue so it has been removed from the dataset.

To study the relationship between mice and gene expression levels, we bicluster the 39×17864 residual matrix computed from the least squares fit to the multiple linear regression model

$$Y_{ij} = \beta_{0j} + \beta_{1j}A_i + \beta_{2j}S_i + \varepsilon_{ij},$$

where Y_{ij} is the log-activation of gene j in mouse i , A_i is the age of mouse i , S_i indicates if mouse i is male, ε_{ij} is a random error, and $(\beta_{0j}, \beta_{1j}, \beta_{2j})$ is a gene-specific coefficient vector.

The entries of the residual matrix are not independent (for example, the sum of each column is zero). Also, the responses of many genes are likely correlated with each other. Thus, the block model required by Theorem 3.1 is not in force, so its conclusion will not hold unless the dependence between the residuals is negligible. In light of this caveat, the example should be considered as exploratory data analysis.

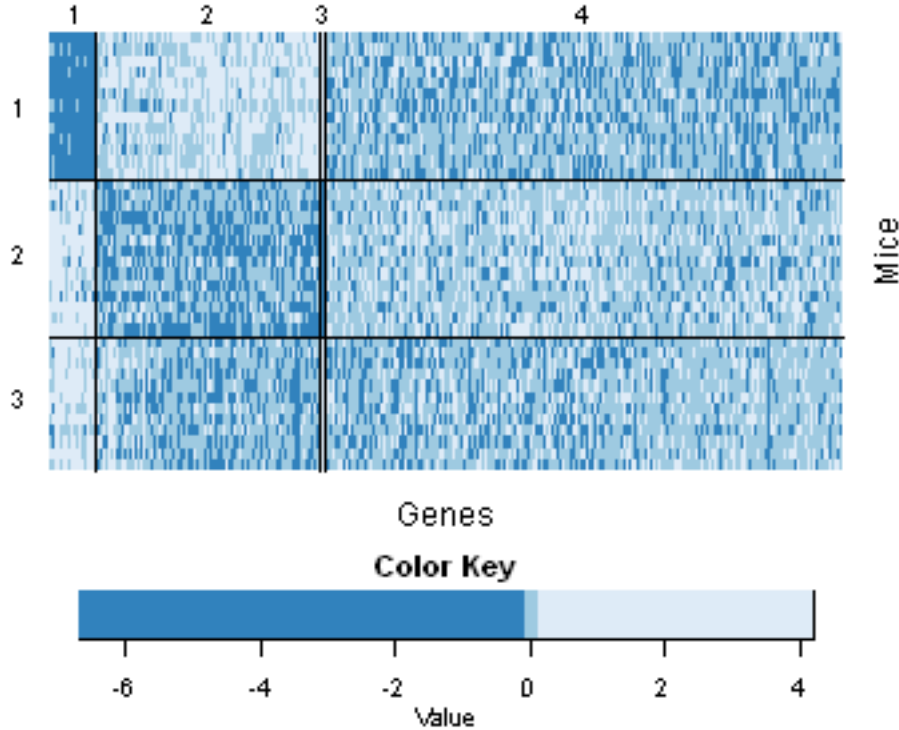
We perform biclustering using profile log-likelihood based on the Gaussian criterion (5.1c). Based on the preliminary analysis in Perry and Owen (2010), it appears that there are three mice groups. To determine an appropriate number of gene groups we experiment with values between two and five. Using four gene groups appears to give a reasonable representation of the data. The heatmap presented in Figure 6.2 shows the results.

Although the expression levels for the fourth gene group appear to be fairly neutral across the three mouse groups, the first three gene groups each have a visually apparent pattern. It appears that a mouse can have high expression levels at most two of the first three gene groups. Mouse group 1 has high expression for gene groups 2 and 3; mouse group 2 has high expression for gene group 1; and mouse group 3 has high expression for gene groups 1 and 3.

The three clusters of mice agree with those found by Perry and Owen (2010). That analysis identified the mouse clusters, but could not attribute

meaning to them. The bicluster-based analysis has deepened our understanding of the three mouse clusters while suggesting some interesting interactions between the genes.

FIG 4. Heatmap generated from AGEMAP residual data reflecting the varying expression patterns in the different biclusters. The colors for the matrix entries correspond encode to the first quartile, the middle two quartiles, and the upper quartile.



7. Discussion. In this report we developed a statistical setting for studying the performance of biclustering algorithms. Under the assumption that the data follows a stochastic block model, we derived sufficient conditions for an algorithm based on maximizing the profile-likelihood to be consistent. This is the first theoretical guarantee for any biclustering algorithm which can be applied to a broad range of data distributions and can handle both sparse and dense data matrices. Our empirical comparisons demonstrated that the method performs well in a variety of situations and can outperform existing procedures.

It is important to note that the theoretical results operate under the as-

sumption that the true number of row and column classes are known. In practice, this is a simplifying assumption and the optimal number of biclusters needs to be inferred from the data. While this is an open problem in the biclustering literature, the formalization of biclustering as an estimation process may help identify parallels that can be drawn from classical model selection procedures to this setting and is an interesting topic for future research.

Applying the profile-likelihood based biclustering algorithm to real data revealed several interesting findings. Our results from the MovieLens dataset identified a distinct difference in movie review behavior and split individuals who only rate recent movies from individuals who rate movies from all time periods. Biclustering the genes and mice in the AGEMAP data exposed an interesting pattern in the expression of certain genes and we found that at most two gene groups can have high expression levels for any one mouse. The consistency theorem proved in this report gives conditions under which we can have confidence in the robustness of these findings.

APPENDIX A: ADDITIONAL TECHNICAL RESULTS

LEMMA A.1. *For each n , let $X_{n,m}$, $1 \leq m \leq n$, be independent random variables with $\mathbb{E} X_{n,m} = 0$. Let ρ_n be a sequence of positive numbers. Let \mathcal{I}_n be a subset of the powerset $2^{[n]}$, with $\inf\{|I| : I \in \mathcal{I}_n\} \geq L_n$. Suppose*

- (i) $\frac{1}{n\rho_n} \sum_{m=1}^n \mathbb{E}|X_{n,m}|^2$ is uniformly bounded in n ;
- (ii) For all $\varepsilon > 0$, $\frac{1}{n\rho_n^2} \sum_{m=1}^n \mathbb{E}(|X_{n,m}|^2; |X_{n,m}| > \varepsilon\sqrt{n}\rho_n) \rightarrow 0$;
- (iii) $\overline{\lim}_{n \rightarrow \infty} \frac{n}{L_n} < \infty$;
- (iv) $\overline{\lim}_{n \rightarrow \infty} \frac{\log|\mathcal{I}_n|}{\sqrt{n}} < \infty$.
- (v) $\overline{\lim}_{n \rightarrow \infty} \rho_n\sqrt{n} = \infty$.

Then

$$\sup_{I \in \mathcal{I}_n} \left| \frac{1}{\rho_n|I|} \sum_{m \in I} X_{n,m} \right| \xrightarrow{P} 0.$$

PROOF. Let $\varepsilon > 0$ be arbitrary. Define $Y_{n,m} = \rho_n^{-1} X_{n,m} \mathbf{I}(|X_{n,m}| \leq \varepsilon\sqrt{n}\rho_n)$, and note that

$$\begin{aligned} \Pr(Y_{n,m} \neq \rho_n^{-1} X_{n,m} \text{ for some } 1 \leq m \leq n) &\leq \sum_{m=1}^n \Pr(|X_{n,m}| > \varepsilon\sqrt{n}\rho_n) \\ &\leq \frac{1}{\varepsilon^2 n \rho_n^2} \sum_{m=1}^n \mathbb{E}(|X_{n,m}|^2; |X_{n,m}| > \varepsilon\sqrt{n}\rho_n) \\ &\rightarrow 0. \end{aligned}$$

Fix an arbitrary $t > 0$. Set $\mu_{n,m} = \mathbb{E} Y_{n,m}$ and for $I \in \mathcal{I}_n$ define

$$\mu_n(I) = \frac{1}{|I|} \sum_{m \in I} \mu_{n,m}.$$

For $I \in \mathcal{I}_n$, write

$$\Pr \left(\sum_{m \in I} Y_{n,m} > t |I| \right) = \Pr \left(\sum_{m \in I} (Y_{n,m} - \mu_{n,m}) > |I|(t - \mu_n(I)) \right)$$

Note that since $\mathbb{E} X_{n,m} = 0$, it follows that

$$|\mu_{n,m}| = |-\mathbb{E}(\rho_n^{-1} X_{n,m}; |X_{n,m}| > \varepsilon \sqrt{n} \rho_n)| \leq \frac{1}{\varepsilon \sqrt{n} \rho_n} \mathbb{E}(|X_{n,m}|^2; |X_{n,m}| > \varepsilon \sqrt{n} \rho_n).$$

Thus, by (ii) and (iii) we have that $\sup_{I \in \mathcal{I}_n} \{|\mu_n(I)|\} \rightarrow 0$; in particular, for n large enough, $\sup_{I \in \mathcal{I}_n} \{|\mu_n(I)|\} < \frac{t}{2}$. Consequently, for n large enough, uniformly for all I ,

$$\Pr \left(\sum_{m \in I} Y_{n,m} > t |I| \right) \leq \Pr \left(\sum_{m \in I} (Y_{n,m} - \mu_{n,m}) > t |I|/2 \right).$$

Similarly,

$$\Pr \left(\sum_{m \in I} Y_{n,m} < -t |I| \right) \leq \Pr \left(\sum_{m \in I} (Y_{n,m} - \mu_{n,m}) < -t |I|/2 \right).$$

We apply Bernstein's inequality to the bound. Define $\sigma_{n,m}^2 = \mathbb{E}(Y_{n,m} - \mu_{n,m})^2$ and $v_n(I) = \sum_{m \in I} \sigma_{n,m}^2$. Note that $|Y_{n,m} - \mu_{n,m}| \leq 2\varepsilon \sqrt{n}$. By Bernstein's inequality,

$$\Pr \left(\left| \sum_{m \in I} (Y_{n,m} - \mu_{n,m}) \right| > t |I|/2 \right) \leq 2 \exp \left\{ - \frac{t^2 |I|^2 / 8}{v_n(I) + \varepsilon t |I| \sqrt{n} / 3} \right\}.$$

By (i), (iv), and (v), it follows that for n large enough, $v_n(I) < \varepsilon t |I| \sqrt{n} / 3$, so

$$\Pr \left(\left| \sum_{m \in I_n} (Y_{n,m} - \mu_{n,m}) \right| > t |I|/2 \right) \leq 2 \exp \left\{ - \frac{t |I|}{8 \varepsilon \sqrt{n}} \right\}.$$

We use this bound for each I to get the union bound:

$$\Pr \left(\sup_{I \in \mathcal{I}_n} \left| \frac{1}{|I|} \sum_{m \in I} Y_{n,m} \right| > t \right) \leq 2 |\mathcal{I}_n| \exp \left\{ - \frac{t L_n}{8 \varepsilon \sqrt{n}} \right\} = 2 \exp \left\{ \log |\mathcal{I}_n| - \frac{t L_n}{8 \varepsilon \sqrt{n}} \right\}.$$

By (iii) and (iv), it is possible to choose ε such that the right hand side goes to zero. \square

PROOF OF LEMMA 4.1. For all $t > 0$,

$$\begin{aligned} \Pr \left(\sup_{\mathcal{J}_\varepsilon} \|\mathbf{R}(\mathbf{g}, \mathbf{h})\|_\infty > t \right) \\ \leq KL \Pr \left(\sup_{I \in \mathcal{I}_n} \rho^{-1} \left| \sum_{\{i,j\} \in I} (X_{ij} - \mu_{c_i d_j}) \right| > t|I| \right), \end{aligned}$$

where $\mathcal{I}_n \subset 2^{[n]} \times 2^{[m]}$ is the set of all biclusters such that $\hat{p}_k > \varepsilon$ for all k and $\hat{q}_l > \varepsilon$ for all l . Since \mathcal{I}_n is a subset of the powerset $2^{[nm]}$, by Lemma A.1, it follows that

$$\Pr \left(\sup_{\mathcal{J}_\varepsilon} \|\mathbf{R}(\mathbf{g}, \mathbf{h})\|_\infty > t \right) \rightarrow 0.$$

□

PROOF OF LEMMA 4.2. The technical assumptions of f imply that its first derivative is bounded. Therefore, f is locally Lipschitz continuous with Lipschitz constant $c = \sup f'(\mu)$ for μ in a neighborhood of \mathcal{M} and

$$|F(\rho^{-1} \bar{\mathbf{X}}(\mathbf{g}, \mathbf{h}), \hat{\mathbf{p}}(\mathbf{g}), \hat{\mathbf{q}}(\mathbf{h})) - G_{\mathbf{M}_0}(\mathbf{C}, \mathbf{D})| \leq c \|\mathbf{R}(\mathbf{g}, \mathbf{h})\|_\infty.$$

From Lemma 4.1, the righthand side converges to zero almost surely and the result follows. □

LEMMA A.2 (Refined Jensen's Inequality). *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be twice differentiable and let \mathcal{N} be a convex set in \mathbb{R} . If x_1, \dots, x_n are points in \mathcal{N} , and if w_1, \dots, w_n are nonnegative numbers summing to one, then*

$$\sum_{i=1}^n w_i f(x_i) - f(z) \geq \frac{1}{2} \inf_{y \in \mathcal{N}} f''(y) \sum_{i=1}^n w_i (x_i - z)^2,$$

where $z = \sum_{i=1}^n w_i x_i$.

PROOF. Define $\kappa = \inf_{y \in \mathcal{N}} f''(y)$ and use the bound

$$f(x_i) \geq f(z) + f'(z)(x_i - z) + \frac{\kappa}{2}(x_i - z)^2.$$

□

LEMMA A.3. *Let $\mathbf{C} = [C_{ak}] \in \mathbb{R}^{A \times K}$ and $\mathbf{D} = [D_{bl}] \in \mathbb{R}^{B \times L}$ be non-negative matrices whose entries sum to one. Let $\mathbf{M} = [\mu_{ab}] \in \mathbb{R}^{A \times B}$ be a matrix with no two identical columns. Define \mathcal{M} to be the convex hull of the entries of \mathbf{M} and let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a twice differentiable convex function*

with f'' bounded away from zero in \mathcal{M} . Suppose that $[\mathbf{C}\mathbf{1}]_a \geq \eta$ for all a and that $D_{bl}D_{b'l} > \varepsilon$ for some l and $b \neq b'$. There exists a positive constant C depending on \mathbf{M} and f such that

$$\begin{aligned} \sum_{k=1}^K \sum_{l=1}^L [\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l f \left(\frac{[\mathbf{C}^T \mathbf{M} \mathbf{D}]_{kl}}{[\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l} \right) \\ \leq \sum_{a=1}^A \sum_{b=1}^B [\mathbf{C}\mathbf{1}]_a [\mathbf{D}\mathbf{1}]_b f(\mu_{ab}) - \frac{C\eta^2\varepsilon}{K^2}. \end{aligned}$$

PROOF. Let l , and $b \neq b'$ be such that $D_{bl}D_{b'l} > \varepsilon$. Since no two columns of \mathbf{M} are identical, there exists an a such that $\mu_{ab} \neq \mu_{ab'}$. Let k be the index of the largest element in row a of matrix \mathbf{C} ; this element must be at least as large as the mean, i.e.

$$C_{ak} \geq \frac{[\mathbf{C}\mathbf{1}]_a}{K} \geq \frac{\eta}{K}.$$

Let $W = [\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l$; this is nonzero. Now, there exists $\mu_* \in \mathcal{M}$ such that

$$[\mathbf{C}^T \mathbf{M} \mathbf{D}]_{kl} = C_{ak} D_{bl} \mu_{ab} + C_{ak} D_{b'l} \mu_{ab'} + (W - C_{ak} D_{bl} - C_{ak} D_{b'l}) \mu_*.$$

Set $\kappa = \inf_{\mu \in \mathcal{M}} f''(\mu)$ and define $\mathbf{N} = [\nu_{ab}] \in \mathbb{R}^{A \times B}$ with $\nu_{ab} = f(\mu_{ab})$. By Lemma A.2,

$$\begin{aligned} [\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l f \left(\frac{[\mathbf{C}^T \mathbf{M} \mathbf{D}]_{kl}}{[\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l} \right) - [\mathbf{C}^T \mathbf{N} \mathbf{D}]_{kl} &\leq -\frac{\kappa}{2} (\mu_{ab} - \mu_{ab'})^2 \frac{C_{ak}^2 D_{bl} D_{b'l}}{W} \\ &\leq -\frac{\kappa \eta^2 \varepsilon}{2K^2} (\mu_{ab} - \mu_{ab'})^2. \end{aligned}$$

This inequality only holds for one particular choice of k and l ; for other choices, the left hand side is nonpositive by Jensen's inequality. The result of the theorem follows, with C defined by

$$C = \frac{\kappa}{2} \min_{a, b \neq b'} (\mu_{ab} - \mu_{ab'})^2.$$

□

PROOF OF LEMMA 4.3. If $\mathbf{D} \notin \mathcal{Q}$, then for some l and some $b \neq b'$, $D_{bl}D_{b'l} \geq \delta$. By Lemma A.3,

$$\sum_{k,l} [\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l f \left(\frac{[\mathbf{C}^T \mathbf{M}_0 \mathbf{D}]_{kl}}{[\mathbf{C}^T \mathbf{1}]_k [\mathbf{D}^T \mathbf{1}]_l} \right) - \sum_{a,b} [\mathbf{C}\mathbf{1}]_a [\mathbf{D}\mathbf{1}]_b f([\mathbf{M}_0]_{a,b}) \leq -\frac{\kappa_0 \eta^2 \delta}{K^2},$$

where κ_0 depends only on \mathbf{M}_0 and f . Similarly, if $\mathbf{C} \notin \mathcal{P}$, then the right hand side is bounded by $-\kappa_0 \eta^2 \delta / L^2$. The result of the lemma follows with $\kappa = \kappa_0 / \min\{K^2, L^2\}$. \square

APPENDIX B: ADDITIONAL EMPIRICAL RESULTS

This appendix reports additional empirical results for Bernoulli and Gaussian distributed data. Figures 5-6 present the average bicluster misclassification rates for each sample size and Tables 3-4 report the standard deviations for the Bernoulli and Gaussian simulations, respectively. Since the normalization for the DI-SIM algorithm is only specified for non-negative data, the algorithm is run on the un-normalized matrix.

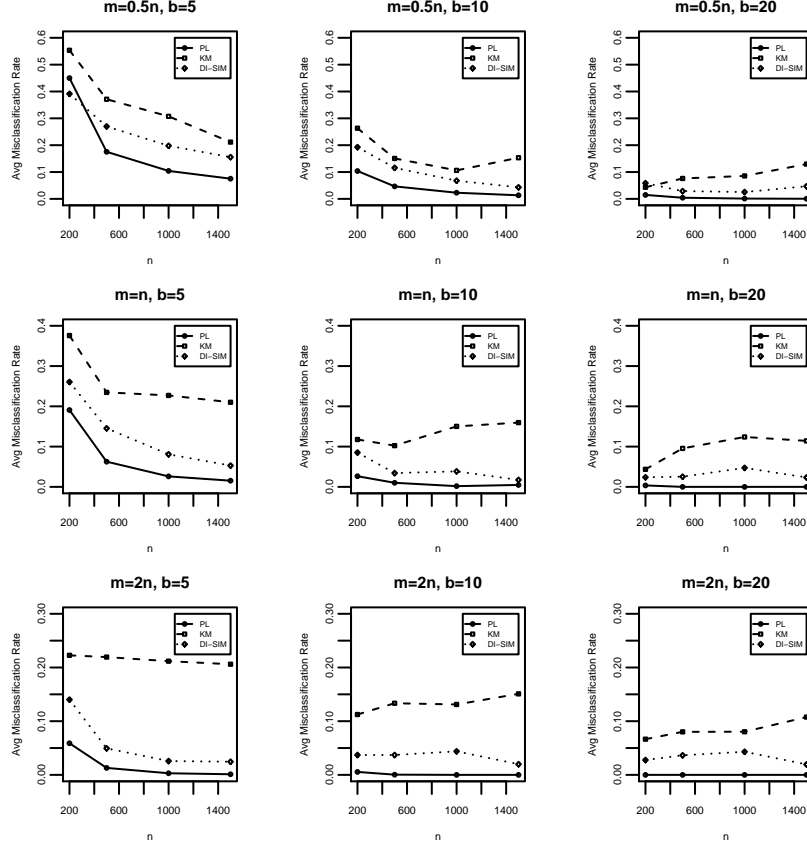
For the Bernoulli simulation, we simulate from a block model with $K = 2$ row clusters and $L = 3$ column clusters. We vary the number of columns, n , between 200 to 1500 and we take the number of rows as $m = \gamma n$ where $\gamma \in \{0.5, 1, 2\}$.

We set the row and column class membership probabilities as $\mathbf{p} = (0.3, 0.7)$ and $\mathbf{q} = (0.2, 0.3, 0.5)$. We choose the matrix of block parameters to be

$$\mathbf{M} = \frac{b}{\sqrt{n}} \begin{pmatrix} 0.43 & 0.06 & 0.13 \\ 0.10 & 0.34 & 0.17 \end{pmatrix}.$$

where the entries were selected to be on the same scale as Bickel and Chen (2009). We vary b between 5 and 20. We generate the data conditional on the row and column classes as $X_{ij} \mid \mathbf{c}, \mathbf{d} \sim \text{Bernoulli}(\mu_{c_i d_j})$.

FIG 5. Average misclassification rates for Bernoulli example over 100 simulations.

TABLE 3
Standard deviations for Bernoulli example over 100 simulations.

$m = 0.5n$								
n	PL			KM			DS	
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.1474	0.0281	0.0344	0.1230	0.1490	0.0982	0.0614	0.0503
500	0.0471	0.0096	0.0029	0.1217	0.1490	0.1399	0.0321	0.0205
1000	0.0111	0.0052	0.0010	0.1573	0.1426	0.1490	0.0223	0.0123
1500	0.0073	0.0030	0.0005	0.1529	0.1720	0.1688	0.0189	0.0087

$m = n$								
n	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.0889	0.0113	0.0285	0.1234	0.1357	0.1112	0.0526	0.0248
500	0.0121	0.0376	0.0003	0.1466	0.1483	0.1544	0.0233	0.0391
1000	0.0056	0.0013	0.0001	0.1659	0.1687	0.1666	0.0132	0.0956
1500	0.0032	0.0403	0.0000	0.1675	0.1720	0.1641	0.0090	0.0637

$m = 2n$								
n	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10
200	0.0275	0.0321	0.0000	0.1336	0.1611	0.1397	0.0373	0.0756
500	0.0048	0.0009	0.0000	0.1598	0.1668	0.1481	0.0120	0.1112
1000	0.0015	0.0001	0.0000	0.1590	0.1652	0.1485	0.0543	0.1230
1500	0.0009	0.0000	0.0000	0.1591	0.1679	0.1617	0.0746	0.0859

For the Gaussian simulation, we simulate from a block model with $K = 2$ row clusters and $L = 3$ column clusters. We vary the number of columns, n , between 50 to 400 and we take the number of rows as $m = \gamma n$ where $\gamma \in \{0.5, 1, 2\}$.

We set the row and column class membership probabilities as $\mathbf{p} = (0.3, 0.7)$ and $\mathbf{q} = (0.2, 0.3, 0.5)$. We choose the matrix of block parameters to be

$$\mathbf{M} = b \begin{pmatrix} 0.47 & 0.15 & -0.60 \\ -0.26 & 0.82 & 0.80 \end{pmatrix}$$

where the entries were simulated from a uniform distribution on $[-1, 1]$. We vary b between 0.5 and 2. We generate the data conditional on the row and column classes as $X_{ij} \mid \mathbf{c}, \mathbf{d} \sim \text{Gaussian}(\mu_{c_id_j}, \sigma = 1)$.

FIG 6. Average misclassification rates for Gaussian example over 500 simulations.

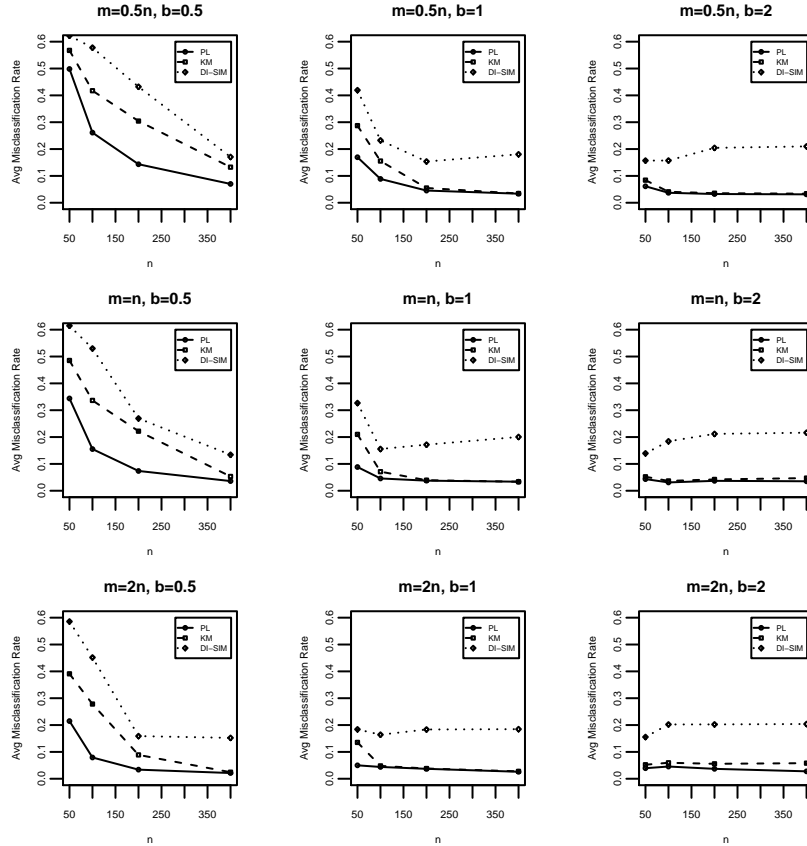


TABLE 4
Standard deviations for Gaussian example over 500 simulations.

$m = 0.5n$									
	PL			KM			DS		
n	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10	b=20
50	0.1300	0.0874	0.1136	0.1019	0.1068	0.1247	0.0961	0.1890	0.1781
100	0.0681	0.0808	0.1094	0.0737	0.1074	0.1116	0.1264	0.1750	0.1975
200	0.0406	0.0958	0.1077	0.0700	0.0966	0.1122	0.1699	0.1846	0.2108
400	0.0459	0.1074	0.1064	0.0754	0.1071	0.1106	0.1274	0.2111	0.2151

$m = n$									
n	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10	b=20
50	0.1121	0.0830	0.1181	0.1010	0.1156	0.1234	0.1076	0.1940	0.1865
100	0.0576	0.0964	0.1047	0.0690	0.1002	0.1121	0.1576	0.1727	0.2060
200	0.0576	0.1120	0.1137	0.0947	0.1119	0.1204	0.1516	0.2033	0.2106
400	0.0843	0.1102	0.1124	0.0819	0.1096	0.1281	0.1712	0.2144	0.2147

$m = 2n$									
n	b=5	b=10	b=20	b=5	b=10	b=20	b=5	b=10	b=20
50	0.0829	0.1014	0.1171	0.0858	0.1290	0.1305	0.1194	0.1604	0.1893
100	0.0527	0.1215	0.1255	0.0851	0.1212	0.1397	0.1696	0.1956	0.2050
200	0.0786	0.1126	0.1126	0.0911	0.1146	0.1364	0.1578	0.2084	0.2106
400	0.0848	0.0966	0.0995	0.0837	0.0992	0.1410	0.2010	0.2118	0.2139

Similar to the Poisson simulation, biclustering based on the profile log-likelihood criterion performs at least as well as the other methods and shows signs of convergence in both examples. These results provide further verification of the theoretical findings and support the use of biclustering based on the profile log-likelihood criterion.

REFERENCES

- ARABIE, P., BOORMAN, S. A. and LEVITT, P. R. (1978). Constructing blockmodels: How and why. *J. Math. Psychol.* **17** 21–63.
- BICKEL, P. J. and CHEN, A. (2009). A Nonparametric View of Network Models and Newman-Girvan and Other Modularities. *Proc. Nat. Acad. Sci. USA* **106** 21068–21073.
- CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. *Proceedings International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8** 93–103.
- CHOI, D., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with growing number of classes. *Biometrika* **99** 273–284.
- DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, Second ed. Springer-Verlag.
- DHILLON, I. S. (2001). Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* 26–29.
- EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA* **95** 14863–14868.
- GETZ, G., LEVINE, E. and DOMANY, E. (2000). Coupled Two-Way Clustering Analysis of Gene Microarray Data. *Proc. Nat. Acad. Sci. USA* **97** 12079–12084.
- GROUPLENS RESEARCH PROJECT, (1998). *MovieLens Dataset*. University of Minnesota <http://www.grouplens.org/data/>.
- HARPAZ, R., PEREZ, H., CHASE, H. S., RABADAN, R., HRIPCSAK, G. and FRIEDMAN, C. (2010). Biclustering of Adverse Drug Events in the FDA’s Spontaneous Reporting System. *Clinical Pharmacology & Therapeutics* **89** 243–250.

- HARTIGAN, J. A. (1972). Direct Clustering of a Data Matrix. *J. Amer. Statist. Assoc.* **67** 123–129.
- HOFMANN, T. (1999). Latent Class Models for Collaborative Filtering. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* 688–693.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic Blockmodels: First Steps. *Social Networks* **5** 109–137.
- KERNIGHAN, B. W. and LIN, S. (1970). An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical journal* **49** 291–307.
- KLUGER, Y., BASRI, R., CHANG, J. T. and GERSTEIN, M. (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research* **13** 703–716.
- LAZZERONI, L. and OWEN, A. (2002). Plaid models for gene expression data. *Statist. Sinica* **12** 61–86.
- MADEIRA, S. C. and OLIVEIRA, A. L. (2004). Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE T. Comput. Bi.* **1** 24–45.
- MIRKIN, B. (1996). *Mathematical classification and clustering*. Kluwer Academic Press.
- MURPHY, S. A. and VAN DER VAART, A. W. (2000). On Profile Likelihood. *J. Amer. Statist. Assoc.* **95** 449–465.
- NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Nat. Acad. Sci. USA* **103** 8577–8582.
- PERRY, P. O. and OWEN, A. B. (2010). A Rotationn Test to Verify Latent Structure. *J. Mach. Learn. Res.* **11** 603–624.
- PERRY, P. O. and WOLFE, P. J. (2012). Null Models for Network Data. Preprint arXiv:1201.5871.
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral Clustering and the High-Dimensional Stochastic Blockmodel. *Ann. Statist.* **39** 1878–1915.
- ROHE, K. and YU, B. (2012). Co-clustering for Directed Graphs; the Stochastic Co-Blockmodel and a Spectral Algorithm. Preprint arXiv:1204.2296.
- TÖSCHER, A., JAHRER, M. and BELL, R. M. (2009). The BigChaos Solution to the Netflix Grand Prize. Technical Report.
- UNGAR, L. and FOSTER, D. P. (1998). A Formal Statistical Approach to Collaborative Filtering. In *In CONALD98*.
- VARADHAN, S. R. S. (2001). *Probability Theory (Courant Lecture Notes)*. American Mathematical Society.
- ZAHN, J. M., POOSALA, S., OWEN, A. B., INGRAM, D. K., LUSTIG, A., CARTER, A., WEERARATNA, A. T., TAUB, D. D., GOROSPE, M., MAZAN-MAMCZARZ, K., LAKATTA, E. G., BOHELER, K. R., XU, X., MATTSON, M. P., FALCO, G., KO, M. S. H., SCHLESSINGER, D., FIRMAN, J., KUMMERFELD, S. K., WOOD, W. H., ZONDERMAN, A. B., KIM, S. K. and BECKER, K. G. (2007). AGEMAP: A Gene Expression Database for Aging in Mice. *PLOS Genetics*.
- ZHAO, Y., LEVINA, E. and ZHU, J. (2011a). Community extraction for social networks. *P. Natl. Acad. Sci. USA* **108** 7321–7326.
- ZHAO, Y., LEVINA, E. and ZHU, J. (2011b). On Consistency of Community Detection in Networks. Preprint arXiv:1110.3854.

LEONARD N. STERN SCHOOL OF BUSINESS
NEW YORK UNIVERSITY
44 WEST 4TH STREET
NEW YORK, NY 10012-1126
E-MAIL: cflynn@stern.nyu.edu
pperry@stern.nyu.edu